GENE EXPRESSION PROFILES IN LIVER CANCER

INVENTORS

Darci Horne, Uwe Scherf and Joseph Vockley

RELATED APPLICATIONS

This application is related to U.S. Provisional Application 60/211,379, filed on June 14, 2000, and U.S. Provisional Application 60/237,054, filed October 2, 2000, which are herein incorporated by reference in their entirety.

BACKGROUND OF THE INVENTION

Primary hepatocellular carcinoma (HCC) is a widespread cancer throughout the world, especially prevalent where the incidence of chronic hepatitis B (HBV) and hepatitis C (HCV) viral infections are endemic (Groen, (1999) Semin. Oncol. Nurs. 15, 48-57; Idilman *et al.*, (1998) J. Viral. Hepat. 5, 110-117; Di Bisceglie *et al.*, (1998) Hepatol. 28, 1161-1165; Johnson, (1997) Hepatogastroenerology 44, 307-312; Sheu, (1997) J. Gastroeneterol. Hepatol. 12, S309-313). Hepatocellular carcinomas are very malignant tumors that generally offer a poor prognosis, dependent on the size of the tumor, the effect on normal liver functions, and the involvement of metastases. They are best treated by surgical resection, when the tumors are diagnosed at a stage where this is a viable possibility, but the recurrence rate for these cancers remains high (Johnson, (1997) Hepatogastroenterology 44, 307-312; Schafer & Sorrell, (1999) Lancet 353, 1253-1257; Groen, (1999) Semin. Oncol. Nurs. 15, 48-57; Sitzman, (1995) World. J. Surg. 19, 790-794; DiCarlo, (1995) Hepato-Gastroenterol. 42, 222-259; Tanaka *et al.*, (1996) Hepato-Gastroenterol. 43, 1172-1181; El-Assal *et al.*, (1997) Surgery 122, 571-577).

Numerous risk factors for the development of HCC have been identified: cirrhosis, HBV or HCV infection, being male, alcohol-related liver disease, exposure to aflatoxins, vinyl chloride and radioactive thorium dioxide, cigarette smoking, ingestion of inorganic arsenic, the use of oral contraceptives and anabolic steroids, iron accumulation, and various inherited metabolic disorders (hemochromatosis, glycogen storage disease, porphyria, tyrosinemia, α-1-antitrypsin deficiency) (Di Bisceglie *et al.*, (1998) Hepatol. 28, 1161-1165; Chen *et al.*, (1997) J. Gastroenterol. Hepatol. 12, S294-308; Schafer & Sorrell (1999) Lancet 353, 1253-1257; Groen, (1999) Semin. Oncol. Nurs. 15, 48-57; Idilman *et al.*, (1998) J. Viral. Hepat. 5, 110-117; Johnson, (1997) Hepato-Gastroenterol. 44, 307-312).

In addition to liver tumors attributed to hepatocellular carcinoma, there are liver tumors that arise as metastases from primary tumors in other parts of the body. These tumors most often metastasize from the gastrointestinal organs, primarily the colon and rectum, but it is possible for metastatic liver cancers to occur from primary cancers throughout the body (Sitzman 1990, Groen 1999). These cancers can be treated using the routine therapies such as chemotherapy, radiotherapy, surgical resection, liver transplantation, chemoembolization, cryosurgery, or a combination of therapies (Sitzman, (1990) Hepatic Neoplasia, in Bayless (editor) Current Therapy in Gastroenterology and Liver Disease, Marcel Dekker; Groen, (1999) Semin. Oncol. Nurs. 15, 48-57).

The characterization of genes that are differentially expressed in tumorigenesis is an important step in identifying those that are intimately involved in the details of a cell's transformation from normal to cancerous. Studies examining the gene expression of metastatic liver tumors and hepatocellular carcinomas in comparison with a set of normal liver tissues would produce data identifying genes that are not expressed in normal livers but have been switched on in tumors, as well as genes that have been completely turned off in these tumors during the progression from a normal to a malignant state. Such studies would also lead to the identification of genes that are expressed in tumor tissue at differing levels, but not expressed at any level in normal liver tissue. The identification of genes and ESTs that are expressed in both types of tumors, *i. e.*, primary hepatocellular carcinomas as well as metastatic tumors of a different origin, and not in normal liver cells would be extremely valuable for the diagnosis of liver cancer.

SUMMARY OF THE INVENTION

The present invention identifies the global changes in gene expression associated with liver cancer by examining gene expression in tissue from normal liver, metastatic malignant liver and hepatocellular carcinoma. The present invention also identifies expression profiles which serve as useful diagnostic markers as well as markers that can be used to monitor disease states, disease progression, drug toxicity, drug efficacy and drug metabolism.

The invention includes methods of diagnosing the presence or absence of liver cancer in a patient comprising the step of detecting the level of expression in a tissue sample of two or more genes from Tables 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of liver cancer. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5. In some preferred embodiments, the method may include detecting the expression level of one or more genes selected from a group consisting Tetraspan NET-6 protein; collagen, type V, alpha; and glypican 3.

The invention also includes methods of detecting the progression of liver cancer and/or differentiating nonmetastatic from metastatic disease. For instance, methods of the invention include detecting the progression of liver cancer in a patient comprising the step of detecting the level of expression in a tissue sample of two or more genes from 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of liver cancer progression. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

In some aspects, the present invention provides a method of monitoring the treatment of a patient with liver cancer, comprising administering a pharmaceutical composition to the patient and preparing a gene expression profile from a cell or tissue sample from the patient and comparing the patient gene expression profile to a gene expression from a cell population comprising normal liver cells or to a gene expression profile from a cell population comprising liver cancer cells or to both. In some preferred embodiments, the gene profile will include the expression level of one or more genes in Tables 3-9. In other preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

PATENT
ATTORNEY DOCKET NO. 44921-5028-US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Darcie T. HORNE *et al.*

Application No.:
(based on 60/237,054)

Filed: June 14, 2001

For: GENE EXPRESSION PROFILES IN LIVER CANCER

Group Art Unit: Unassigned

Examiner: Unassigned

Commissioner for Patents
Washington, D.C. 20231
BOX SEQUENCE

STATEMENT ACCOMPANYING SEQUENCE LISTING

Dear Sir:

The undersigned hereby states upon information and belief that the Sequence Listing submitted concurrently herewith does not include matter which goes beyond the content of the application as filed and that the information recorded on the diskette submitted concurrently herewith is identical to the written Sequence Listing submitted herewith.

Respectfully submitted,
MORGAN, LEWIS & BOCKIUS LLP

Dated: 6/14/01

By: Rosanne Kosson
Rosanne Kosson
Reg. No. 46,840

Customer No. 009629
MORGAN, LEWIS & BOCKIUS LLP
1800 M Street, NW
Washington, D.C. 20036
Tel: 202-467-7000
Fax: 202-467-7258

In another aspect, the present invention provides a method of treating a patient with liver cancer, comprising administering to the patient a pharmaceutical composition, wherein the composition alters the expression of at least one gene in Tables 3-9, preparing a gene expression profile from a cell or tissue sample from the patient comprising tumor cells and comparing the patient expression profile to a gene expression profile from an untreated cell population comprising liver cancer cells. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

In one aspect, the present invention provides a method of diagnosing hepatocellular carcinoma in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9, wherein differential expression of the genes in Tables 3-9 is indicative of hepatocellular carcinoma. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5

In another aspect, the present invention provides a method of detecting the progression of hepatocellular carcinoma in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of hepatocellular carcinoma progression. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5.

The present invention also provides materials and methods for monitoring the treatment of a patient with a hepatocellular caricnoma. The present invention provides a method of monitoring the treatment of a patient with hepatocellular carcinoma, comprising administering a pharmaceutical composition to the patient, preparing a gene expression profile from a cell or tissue sample from the patient and comparing the patient gene expression profile to a gene expression from a cell population comprising normal liver cells or to a gene expression profile from a cell population comprising hepatocellular carcinoma cells or to both. In some preferred embodiments, the method may include detecting the level of expression of one or more genes from the genes listed in Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5.

In a related aspect, the present invention provides a method of treating a patient with hepatocellular carcinoma, comprising administering to the patient a pharmaceutical composition, wherein the composition alters the expression of at least one gene in Tables 3-

9, preparing a gene expression profile from a cell or tissue sample from the patient comprising hepatocellular carcinoma cells and comparing the patient expression profile to a gene expression profile from an untreated cell population comprising hepatocellular carcinoma cells. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5.

The present invention provides a method of diagnosing a metastatic liver tumor in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9, wherein differential expression of the genes in Tables 3-9 is indicative of hepatocellular carcinoma. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

The present invention provides a method of detecting the progression of a metastatic liver tumor in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9, wherein differential expression of the genes in Tables 3-9 is indicative of a metastatic liver tumor progression. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

In a related aspect, the present invention provides a method of monitoring the treatment of a patient with a metastatic liver tumor, comprising administering a pharmaceutical composition to the patient, preparing a gene expression profile from a cell or tissue sample from the patient and comparing the patient gene expression profile to a gene expression from a cell population comprising normal liver cells or to a gene expression profile from a cell population comprising metastatic liver tumor cells or to both. In some preferred embodiments, the method of the present invention may include detecting the expression level of one or more genes selected from the genes listed in Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

In some preferred embodiments, the present invention provides a method of treating a patient with a metastatic liver tumor, comprising administering to the patient a pharmaceutical composition, wherein the composition alters the expression of at least one gene in Tables 3-9, preparing a gene expression profile from a cell or tissue sample from the patient comprising metastatic liver tumor cells and comparing the patient expression profile to a gene expression profile from an untreated cell population comprising metastatic liver tumor cells. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

The invention also includes methods of differentiating metastatic liver cancer from hepatocellular carcinoma in a patient comprising the step of detecting the level of expression in a tissue sample of two or more genes from Tables 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of metastatic liver cancer rather than hepatocellular carcinoma.

The invention further includes methods of screening for an agent capable of modulating the onset or progression of liver cancer, comprising the steps of exposing a cell to the agent; and detecting the expression level of two or more genes from Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

Any of the methods of the invention described above may include the detection of at least 2 genes from the tables. Preferred methods may detect all or nearly all of the genes in the tables. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

The invention further includes compositions comprising at least two oligonucleotides, wherein each of the oligonucleotides comprises a sequence that specifically hybridizes to a gene in Tables 3-9 as well as solid supports comprising at least two probes, wherein each of the probes comprises a sequence that specifically hybridizes to a gene in Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

The invention further includes computer systems comprising a database containing information identifying the expression level in liver tissue of a set of genes comprising at least two genes in Tables 3-9; and a user interface to view the information. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5. The database may further include sequence information for the genes, information identifying the expression level for the set of genes in normal liver tissue and malignant tissue (metastatic and nonmetastatic) and may contain links to external databases such as GenBank.

The invention further comprises kits useful for the practice of one or more of the methods of the invention. In some preferred embodiments, a kit may contain one or more solid supports having attached thereto one or more oligonucleotides. The solid support may be a high-density oligonucleotide array. Kits may further comprise one or more reagents for use with the arrays, one or more signal detection and/or array-processing instruments, one or more gene expression databases and one or more analysis and database management software packages.

Lastly, the invention includes methods of using the databases, such as methods of using the disclosed computer systems to present information identifying the expression level in a tissue or cell of at least one gene in Tables 3-9, comprising the step of comparing the expression level of at least one gene in Tables 3-9 in the tissue or cell to the level of expression of the gene in the database. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a flow chart showing a schematic representation of the experimental protocol.

Figures 2A-2C are graphs of the number of genes present in all samples as a function of the number of samples for the second sample set. Figure 2A is the Gene Signature Curve for normal liver tissue. Figure 2B is the Gene Signature Curve for metastatic liver tumor samples. Figure 2C is the Gene Signature Curve for hepatocellular carinoma samples.

DETAILED DESCRIPTION

Many biological functions are accomplished by altering the expression of various genes through transcriptional (*e.g.*, through control of initiation, provision of RNA precursors, RNA processing, etc.) and/or translational control. For example, fundamental biological processes such as cell cycle, cell differentiation and cell death, are often characterized by the variations in the expression levels of groups of genes.

Changes in gene expression also are associated with pathogenesis. For example, the lack of sufficient expression of functional tumor suppressor genes and/or the over expression of oncogene/protooncogenes could lead to tumorgenesis or hyperplastic growth of cells (Marshall, (1991) Cell, 64, 313-326; Weinberg, (1991) Science, 254, 1138-1146). Thus, changes in the expression levels of particular genes (*e.g.*, oncogenes or tumor suppressors) serve as signposts for the presence and progression of various diseases.

Monitoring changes in gene expression may also provide certain advantages during drug screening development. Often drugs are screened and prescreened for the ability to interact with a major target without regard to other effects the drugs have on cells. Often such other effects cause toxicity in the whole animal, which prevent the development and use of the potential drug.

The present inventors have examined tissue samples from normal liver, metastatic malignant liver and hepatocellular carcinoma to identify the global changes in gene expression associated with liver cancer. The protocol used is schematically represented in Figure 1. These global changes in gene expression, also referred to as expression profiles, provide useful markers for diagnostic uses as well as markers that can be used to monitor disease states, disease progression, drug toxicity, drug efficacy and drug metabolism.

The present invention provides compositions and methods to detect the level of expression of genes that may be differentially expressed dependent upon the state of the cell, *i.e.*, normal versus cancerous. As used herein, the phrase "detecting the level expression" includes methods that quantitate expression levels as well as methods that determine whether a gene of interest is expressed at all. Thus, an assay which provides a yes or no result without necessarily providing quantification of an amount of expression is an assay that requires "detecting the level of expression" as that phrase is used herein.

Assay Formats

The genes identified as being differentially expressed in liver cancer may be used in a variety of nucleic acid detection assays to detect or quantititate the expression level of a gene or multiple genes in a given sample. For example, traditional Northern blotting, nuclease protection, RT-PCR and differential display methods may be used for detecting gene expression levels. Those methods are useful for some embodiments of the invention. However, methods and assays of the invention are most efficiently designed with array or chip hybridization-based methods for detecting the expression of a large number of genes.

Any hybridization assay format may be used, including solution-based and solid support-based assay formats. Solid supports containing oligonucleotide probes for differentially expressed genes of the invention can be filters, polyvinyl chloride dishes, silicon or glass based chips, etc. Such wafers and hybridization methods are widely available, for example, those disclosed by Beattie (WO 95/11755). Any solid surface to which oligonucleotides can be bound, either directly or indirectly, either covalently or non-covalently, can be used. A preferred solid support is a high density array or DNA chip. These contain a particular oligonucleotide probe in a predetermined location on the array. Each predetermined location may contain more than one molecule of the probe, but each molecule within the predetermined location has an identical sequence. Such predetermined locations are termed features. There may be, for example, about 2, 10, 100, 1000 to 10,000; 100,000 or 400,000 of such features on a single solid support. The solid support, or the area within which the probes are attached may be on the order of a square centimeter.

Oligonucleotide probe arrays for expression monitoring can be made and used according to any techniques known in the art (see for example, Lockhart *et al.*, (1996) Nat. Biotechnol. 14, 1675-1680; McGall *et al.*, (1996) Proc. Nat. Acad. Sci. USA 93, 13555-13460). Such probe arrays may contain at least two or more oligonucleotides that are complementary to or hybridize to two or more of the genes described herein. Such arrays may also contain oligonucleotides that are complementary or hybridize to at least about 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 50, 70, 100 or or more the genes described herein.

The genes which are assayed according to the present invention are typically in the form of mRNA or reverse transcribed mRNA. The genes may be cloned or not and the genes may be amplified or not. The cloning itself does not appear to bias the representation of genes within a population. However, it may be preferable to use polyA+ RNA as a source, as it can be used with less processing steps.

The sequences of the expression marker genes are in the public databases. Tables 3-9 provide the GenBank accession number for the genes and ESTs identified called either Accession # (Tables 3, 4, and 5) or Fragment Name (Tables 6-9). The sequences of the genes in GenBank are expressly incorporated by reference as are equivalent and related sequences present in GenBank or other public databases. The column labeled "SEQ ID" refers to the sequence identification number correlating the listed gene or EST to its sequence information as provided within the sequence listing of this application.

Probes based on the sequences of the genes described herein may be prepared by any commonly available method. Oligonucleotide probes for assaying the tissue or cell sample are preferably of sufficient length to specifically hybridize only to appropriate, complementary genes or transcripts. Typically the oligonucleotide probes will be at least 10, 12, 14, 16, 18, 20 or 25 nucleotides in length. In some cases longer probes of at least 30, 40, or 50 nucleotides will be desirable.

As used herein, oligonucleotide sequences that are complementary to one or more of the genes described herein, refers to oligonucleotides that are capable of hybridizing under stringent conditions to at least part of the nucleotide sequence of said genes. Such hybridizable oligonucleotides will typically exhibit at least about 75% sequence identity at the nucleotide level to said genes, preferably about 80% or 85% sequence identity or more preferably about 90% or 95% or more sequence identity to said genes.

"Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding, or other interactions, between the labeled target nucleic acids and components of the oligonucleotide array (*e.g.*, the oligonucleotide probes, control probes, the array substrate, etc.). Background signals may also be produced by intrinsic fluorescence of the array components themselves. A single background signal can be calculated for the entire array, or a different background signal may be calculated for each target nucleic acid. In a preferred embodiment, background is calculated as the average hybridization signal intensity for the lowest 5% to 10% of the probes in the array, or, where a different background signal is calculated for each target gene, for the lowest 5% to 10% of the probes for each gene. Of course, one of skill in the art will appreciate that where the probes to a particular gene hybridize well and thus appear to be specifically binding to a target sequence, they should not be used in a background signal calculation. Alternatively, background may be calculated as the average hybridization signal intensity produced by hybridization to probes that are not complementary to any sequence found in the sample (*e.g.*, probes directed to nucleic acids of the opposite sense or to genes not found in the sample such as bacterial genes where the sample is mammalian nucleic acids). Background can also be calculated as the average signal intensity produced by regions of the array that lack any probes at all.

The phrase "hybridizing specifically to" refers to the binding, duplexing or hybridizing of a molecule substantially to or only to a particular nucleotide sequence or sequences under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA.

Assays and methods of the invention may utilize available formats to simultaneously screen at least about 100, preferably about 1000, more preferably about 10,000 and most preferably about 1,000,000 or more different nucleic acid hybridizations.

The term "mismatch control" or "mismatch probe" refer to a probe whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in a high-density array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence. The mismatch may comprise one or more bases.

While the mismatch(s) may be located anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The perfect match (PM) probe can be a "test probe", a "normalization control" probe, an expression level control probe and the like. A perfect match control or perfect match probe is, however, distinguished from a "mismatch control" or "mismatch probe."

As used herein a "probe" is defined as a nucleic acid, capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, a probe may include natural (*i.e.*, A, G, U, C or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in probes may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages.

The term "stringent conditions" refers to conditions under which a probe will hybridize to its target subsequence, but with only insubstantial hybridization to other sequences or to other sequences such that the difference may be identified. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH.

Typically, stringent conditions will be those in which the salt concentration is at least about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50 nucleotide). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

The "percentage of sequence identity" or "sequence identity" is determined by comparing two optimally aligned sequences or subsequences over a comparison window or span, wherein the portion of the polynucleotide sequence in the comparison window may optionally comprise additions or deletions (*i.e.*, gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical monomer unit (*e.g.*, nucleic acid base or amino acid residue) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Percentage sequence identity when calculated using the programs GAP or BESTFIT (see below) is calculated using default gap weights.

Homology or identity may be determined by **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) analysis using the algorithm employed by the programs **blastp**, **blastn**, **blastx**, **tblastn** and **tblastx** (Karlin *et al.*, (1990) Proc. Natl. Acad. Sci. USA 87, 2264-2268 and Altschul, (1993) J. Mol. Evol. 36, 290-300, fully incorporated by reference) which are tailored for sequence similarity searching. The approach used by the **BLAST** program is to first consider similar segments between a query sequence and a database sequence, then to evaluate the statistical significance of all matches that are identified and finally to summarize only those matches which satisfy a preselected threshold of significance. For a discussion of basic issues in similarity searching of sequence databases, see Altschul *et al.*, (1994) Nature Genet. 6, 119-129) which is fully incorporated by reference. The search parameters for **histogram**, **descriptions**, **alignments**, **expect** (*i.e.*, the statistical significance threshold for reporting matches against database sequences), **cutoff, matrix** and **filter** are at the default settings. The default scoring matrix used by **blastp, blastx, tblastn**, and **tblastx** is the **BLOSUM62** matrix (Henikoff *et al.*, (1992) Proc. Natl. Acad. Sci. USA 89, 10915-10919, fully incorporated by reference). Four **blastn** parameters were adjusted as follows: Q=10 (gap creation penalty); R=10 (gap extension penalty); wink=1 (generates word hits at every winkth position along the query); and gapw=16 (sets the window width within which gapped alignments are generated). The equivalent **Blastp** parameter settings were Q=9; R=2; wink=1; and gapw=32. A **Bestfit** comparison between sequences, available in the GCG package version 10.0, uses DNA parameters GAP=50 (gap creation penalty) and LEN=3 (gap extension penalty) and the equivalent settings in protein comparisons are GAP=8 and LEN=2.

Probe design

One of skill in the art will appreciate that an enormous number of array designs are suitable for the practice of this invention. The high density array will typically include a number of probes that specifically hybridize to the sequences of interest. See WO 99/32660 for methods of producing probes for a given gene or genes. In addition, in a preferred embodiment, the array will include one or more control probes.

High density array chips of the invention include "test probes." Test probes may be oligonucleotides that range from about 5 to about 500 or about 5 to about 50 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In other particularly preferred embodiments the probes are about 20 to 25 nucleotides in length. In another preferred embodiment, test probes are double or single strand DNA sequences. DNA sequences are isolated or cloned from natural sources or amplified from natural sources using natural nucleic acid as templates. These probes have sequences complementary to particular subsequences of the genes whose expression they are designed to detect. Thus, the test probes are capable of specifically hybridizing to the target nucleic acid they are to detect.

In addition to test probes that bind the target nucleic acid(s) of interest, the high density array can contain a number of control probes. The control probes fall into three categories referred to herein as (1) normalization controls; (2) expression level controls; and (3) mismatch controls.

Normalization controls are oligonucleotide or other nucleic acid probes that are complementary to labeled reference oligonucleotides or other nucleic acid sequences that are added to the nucleic acid sample. The signals obtained from the normalization controls after hybridization provide a control for variations in hybridization conditions, label intensity, "reading" efficiency and other factors that may cause the signal of a perfect hybridization to vary between arrays. In a preferred embodiment, signals (*e.g.*, fluorescence intensity) read from all other probes in the array are divided by the signal (*e.g.*, fluorescence intensity) from the control probes thereby normalizing the measurements.

Virtually any probe may serve as a normalization control. However, it is recognized that hybridization efficiency varies with base composition and probe length. Preferred normalization probes are selected to reflect the average length of the other probes present in the array, however, they can be selected to cover a range of lengths. The normalization control(s) can also be selected to reflect the (average) base composition of the other probes in the array, however in a preferred embodiment, only one or a few probes are used and they are selected such that they hybridize well (*i.e.*, no secondary structure) and do not match any target-specific probes.

Expression level controls are probes that hybridize specifically with constitutively expressed genes in the biological sample. Virtually any constitutively expressed gene provides a suitable target for expression level controls. Typical expression level control probes have sequences complementary to subsequences of constitutively expressed "housekeeping genes" including, but not limited to the β-actin gene, the transferrin receptor gene, the GAPDH gene, and the like.

Mismatch controls may also be provided for the probes to the target genes, for expression level controls or for normalization controls. Mismatch controls are oligonucleotide probes or other nucleic acid probes identical to their corresponding test or control probes except for the presence of one or more mismatched bases. A mismatched base is a base selected so that it is not complementary to the corresponding base in the target sequence to which the probe would otherwise specifically hybridize. One or more mismatches are selected such that under appropriate hybridization conditions (*e.g.*, stringent conditions) the test or control probe would be expected to hybridize with its target sequence, but the mismatch probe would not hybridize (or would hybridize to a significantly lesser extent). Preferred mismatch probes contain a central mismatch. Thus, for example, where a probe is a twenty-mer, a corresponding mismatch probe will have the identical sequence except for a single base mismatch (*e.g.*, substituting a G, a C or a T for an A) at any of positions 6 through 14 (the central mismatch).

Mismatch probes thus provide a control for non-specific binding or cross hybridization to a nucleic acid in the sample other than the target to which the probe is directed. Mismatch probes also indicate whether a hybridization is specific or not. For example, if the target is present the perfect match probes should be consistently brighter than the mismatch probes. In addition, if all central mismatches are present, the mismatch probes can be used to detect a mutation. The difference in intensity between the perfect match and the mismatch probe ($I_{(PM)} - I_{(MM)}$) provides a good measure of the concentration of the hybridized material.

Nucleic Acid Samples

As is apparent to one of ordinary skill in the art, nucleic acid samples used in the methods and assays of the invention may be prepared by any available method or process. Methods of isolating total mRNA are also well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in detail in Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I Theory and Nucleic Acid Preparation, Tijssen, (1993) (editor) Elsevier Press. Such samples include RNA samples, but also include cDNA synthesized from a mRNA sample isolated from a cell or tissue of interest. Such samples also include DNA amplified from the cDNA, and an RNA transcribed from the amplified DNA. One of skill in the art would appreciate that it is desirable to inhibit or destroy RNase present in homogenates before homogenates can be used.

Biological samples may be of any biological tissue or fluid or cells from any organism as well as cells raised *in vitro*, such as cell lines and tissue culture cells. Frequently the sample will be a “clinical sample” which is a sample derived from a patient. Typical clinical samples include, but are not limited to, sputum, blood, blood-cells (*e.g.*, white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom.

Biological samples may also include sections of tissues, such as frozen sections or formalin fixed sections taken for histological purposes.